Veterinary Immunology and Immunopathology 138 (2010) 241–251

Contents lists available at ScienceDirect

Veterinary Immunology and Immunopathology

journal homepage: www.elsevier.com/locate/vetimm

Utility, limitations, and promise of proteomics in animal science

John D. Lippolis*, Timothy A. Reinhardt

Ruminant Diseases and Immunology Research Unit, USDA-ARS, National Animal Disease Center, 1920 Dayton Avenue, Ames, IA 50010, United States

ARTICLE INFO

Keywords:
Proteomics
Mass spectrometry
Protein expression

ABSTRACT

Proteomics experiments have the ability to simultaneously identify and quantify thousands of proteins in one experiment. The use of this technology in veterinary/animal science is still in its infancy, yet it holds significant promise as a method for advancing veterinary/animal science research. Examples of current experimental designs and capabilities of proteomic technology and basic principles of mass spectrometry are discussed. In addition, challenges and limitations of proteomics are presented, stressing those that are unique to veterinary/animal sciences.

Published by Elsevier B.V.

1. Background: transition from genomic studies to proteomic studies

Proteomics has rapidly moved from a relatively new technology to a rapidly maturing essential tool in the omics age. Its existence is largely due to the success of the genome projects as well as rapid advancements in commercial mass spectrometers. The field of proteomics could not exist without the success of the genome projects. The genome projects of the various domestic animals will continue to increase our ability to associate desired traits with the necessary genes/proteins. Genome projects have given us the gene sequence and gene expression information with the use of techniques such as real-time PCR and microarray assays. However, understanding the genes is only the beginning of the biological story. Proteomics is giving us information regarding protein expression and post-translational modifications. Together, these techniques have brought animal research to a molecular level.

Proteomics is the large-scale study of protein expression, protein–protein interactions, or post-translational modifications (for more specific reviews see Cravatt et al., 2007; Gingras et al., 2007; Ong and Mann, 2005; Witze et al., 2007). Unlike other methodologies that analyze a few proteins at a time, proteomics can analyze thousands of proteins in a single experiment. This ability to analyze thousands of proteins gives the field of proteomics a unique capability to demonstrate how cells dynamically respond to changes in their environment. Therefore, a goal of proteomics is to identify new and potentially unexpected changes in protein expression, interaction or modification as a result of an experimental treatment. Generation of large proteomic data sets is expected to demonstrate the interdependence of cellular processes important for normal cell growth or a cell's response to abnormal or disease conditions. In essence, a proteomic approach enables an investigator to step back and without prejudice, view the whole picture of cellular functions instead of one particular action of one protein. This type of research enables the discovery of unexpected connections between cellular processes and can serve as a precursor to new hypotheses.

2. Utility: what questions does a proteomics experiment ask?

Proteins play many fundamental roles in all biological processes. Some functions of proteins include: structural building blocks, conduits of information, controllers of chemical reactions, and antimicrobial defense mechanisms. The functional abilities of cells are dynamic as cells respond to stimuli or stresses. Much of a cell's response to stimuli or stress is manifest by the alteration of the

* Corresponding author. Tel.: +1 515 337 7446.
E-mail address: john.lippolis@ars.usda.gov (J.D. Lippolis).

0165-2427/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.vetimm.2010.10.003

expression levels of various proteins. Identification and understanding of proteins involved in biological processes have been goals of scientists for decades. This ability to analyze thousands of proteins gives the field of proteomics a unique capability to demonstrate how cells can dynamically respond to changes in their environment.

Proteomic experiments can be directed towards detection of certain known proteins of interest, or an indirect or shotgun approach can be taken. In this review we will concentrate on the proteomic approach that has the goal of identifying the greatest number of proteins in a biological sample. This type of broad proteomic experiment is referred to as "shotgun" proteomics. The ultimate goal of a shotgun proteomics experiment is to identify all the proteins in a sample. However, the number of proteins in a sample, the dynamic range of expression of the various proteins in a sample, and the limitation of the mass spectrometers, make this a very difficult goal to achieve. Consequently, the preparation of the sample plays a significant role in the amount of protein information one can extract from a proteomic experiment. Currently a single proteomics experiment can yield hundreds, and at times thousands, of protein identifications (Aebersold and Mann, 2003).

Shotgun proteomic experiments can be divided into two groups, survey and expression. A survey proteomics experiment is an experiment designed to identify as many proteins in a sample as possible. Whereas expression proteomics adds quantification of proteins to the goal of identifying as many proteins as possible in a given sample, for the purpose of assessing the potential effects of an experimental treatment.

2.1. *Survey proteomics*

Survey proteomics data allows an investigator to accomplish two basic goals. First, to obtain an overview of the types of proteins identified in a specific cell or tissue, and to then organize them by biological process, cellular compartment, or molecular function. From this type of data general conclusions about the types of proteins that are important for that cell's function can be reached. For example, in a survey proteomics experiment of bovine neutrophils it was found that 25% of the proteins identified fell into two functional categories, immune functions and cellular mobility (Lippolis and Reinhardt, 2005). This observation is in harmony with the known function of neutrophils, which is to relocate themselves to the site of an infection and to release various antimicrobial agents important in the resolution of an infection. Alternatively, a survey proteomic experiment that identified proteins of the bovine milk fat globule membrane showed that nearly 60% of the proteins observed fell into three functional categories, membrane/protein trafficking, cell signaling, and fat transport and metabolism (Reinhardt and Lippolis, 2006). Again, this observation is in harmony with the known functions of the milk fat globule membrane and the secretory mammary epithelial cells (sMEC) from which they are derived. From these data one can quickly observe that a proteome is not a set of conserved proteins with a few proteins that distinguish various samples, but the protein profile observed from a sMEC is quite different from that observed from the membrane of a neutrophil.

A more specific example of the utility of survey proteomics can be illustrated by the following. The proteomic survey of the milk fat globule membrane found the toll-like receptor (TLR) 2 and 4 proteins (Reinhardt and Lippolis, 2006). In this report the authors determine that two TLRs and CD14 were present on milk fat globule membranes and therefore by extension on the apical membrane of sMEC. This observation helps to explain the mechanism of how sMEC respond to the presence of a pathogen and that these cells likely play an important role in the innate immune response.

2.2. *Expression proteomics*

In addition to the more general categorization of proteins, the second basic goal of a survey proteomics is to identify interesting proteins for further study. Expression proteomics uses experimental treatment with the dual goal of identifying proteins and measuring experimentally induced changes in the expression of those proteins. Most often the type of quantitation is relative, meaning that two or more samples are compared and one of the samples acts as a reference to which the others are measured (Bantscheff et al., 2007; Lippolis and Reinhardt, 2008; Ong and Mann, 2005). For example, an experiment that compared the effect of growth of *Escherichia coli* in whole milk compared each protein identified to the same protein from the same *E. coli* grown in laboratory media (Lippolis et al., 2009). The result was information about the relative expression of 1000 proteins and any changes of protein expression due to a change in growth media. It has been shown that growth rates of various strains of *E. coli* in an infected gland prior to the immune response have been correlated with final severity of the infection (Kornalijnslijper et al., 2004). The interrogation of this type of proteomic data can yield important clues about protein function and in this case the proteins involved in the adaptation of *E. coli* for growth in milk and subsequently the bacteria's ability to cause an infection. Of the 1000 proteins that were identified with expression data, 20% were up regulated in *E. coli* grown in milk compared to laboratory media and 8% were down regulated.

One group of proteins that were up regulated in *E. coli* grown in milk was those whose function involved iron transport. Iron is an essential element for the growth of bacteria and the availability of iron has been linked to the pathogenicity of the bacteria (Klebba, 2003). Lactoferrin is a milk protein that binds and is part of the body's innate mechanism to sequester iron to deprive invading pathogens of this necessary nutrient (Legrand et al., 2005). *E. coli* grown in milk were shown to up regulate four outer membrane siderophore receptors (FecA, FepA, FhuA and Fiu), suggesting that the bacteria are reacting to the milk environment by increasing their ability to remove iron from their environment. An attempt has been made to block iron uptake by *E. coli* by generating an immune response to one of the siderophore receptors (FecA). The researchers were successful in generating antibody but unsuccessful in changing the clinical severity of an infection (Takemura et al., 2002). This proteomic data has demonstrated the redundancy siderophore receptors, and may explain the lack of effect on the blocking of only one of a series of siderophore receptors in *E. coli*. The proteomic data further suggests that any successful blocking of iron uptake for the purpose of moderating an infection would require blocking multiple iron acquisition pathways (Lippolis et al., 2009).

3. Limitations: what are some difficulties of a proteomic experiment?

The promise of proteomics does come with a number of difficulties that must be addressed or acknowledged to reduce the limitations of this technology. Some of the factors that limit proteomics are the quality of the genomic databases, the complexity and dynamic range of proteins in a sample, the capabilities of various mass spectrometers, and the cost of these experiments.

3.1. Quality of the genomic database

Proteomics owes its existence, in part, to the Human Genome Project. Completed in 2003, this 13-year project had as two of its goals to identify all of the genes in human DNA and determine the sequence of the 3 billion chemical base pairs that make up human DNA. The success of this project has resulted in sequencing projects in other lab animal models, companion animals, and economically important agricultural animals and plants. However, the genomic databases of these species are not at the same level of completeness. As sequencing projects for these various species are completed the number and confidence of proteomic techniques to identify proteins will be increased.

Many proteomic software packages have been developed that align the data obtained from the mass spectrometer with various protein databases (e.g., NCBI non-redundant protein database, Swissprot). Two of the most common software packages (Mascot, Sequest) assume that the protein database contains all proteins of interest. If this assumption is not true then the software will likely identify a homologous protein from a species that has been more completely sequenced. However, we have found that the protein score will be reduced if mass spectrometry data is not matched to the correct species. A reduced score could lead to a misidentification or a lack of identification. To illustrate this point, the same mass spectrometry dataset was processed using a database from 2004 or 2007. Importantly, in 2006 the Bovine Genome Sequencing Center released an updated version of the genomic assembly, adding more genomic information to the bovine protein database. The data showed that twice as many proteins were identified as bovine using the 2007 database then using the 2004 database. Not only did the number of proteins identified as bovine increase but also the average protein score increased, indicating greater confidence in the identification. From these data we conclude that a more complete database results in a better data set as determined by the number of proteins of the correct species and the higher identification scores (Lippolis and Reinhardt, 2008).

The presence, absence or modification of a protein has limited value without knowledge of the function of the protein. The modulation of a protein with a known function can then be associated with a cellular compartment, biological process, or molecular function. In contrast, the modulation of a protein with an unknown function tells little to nothing about the biological, cellular, or molecular functions of the cell type. The challenge of proteomics is to sort through a mountain of data and find the information about protein changes that are critical to the cells response to changes in its environment. Continued efforts to categorize proteins according to their known or predicted functions are necessary and are currently underway. These efforts are critical to understand all the information that a proteomic experiment can yield. Consortiums such as The Gene Ontology (www.geneontology.org) and the Kyoto Encyclopedia of Genes and Genomes (www.genome.jp/kegg) provide means to group proteins by function or into biological processes.

3.2. Mass spectrometers

The improvement in proteomic analysis is largely due to the advancement in the field of mass spectrometers. Many mass spectrometers can detect and identify peptides in the femtomole (10^{-15}) to attomole (10^{-18}) range (Moyer et al., 2003). In fact, investigators have been able to sequence as low as 10 amol of trypsin-digested cytochrome *c* (Martin et al., 2000). The level of sensitivity possible by some mass spectrometers allows investigators to identify proteins from a relatively small number of cells. For example, 160 fmol of protein is approximately 10^{10} molecules; if that protein were expressed in a cell at 1000 copies per cell, this protein would be detectable from 10^7 cells. This number of cells is easily obtainable from bacterial samples and eukaryotic cells. However, instrument sensitivity is not the only factor that is important to the detection and identification of proteins by mass spectrometry. Resolution, dynamic range, the ability to select and separate an ion of interest, and the ability to trap and store ions are all important factors that affect the utility of mass spectrometry in the field of protein chemistry. There are many types of mass spectrometers that can be used for proteomic studies, and each accomplishes the task of protein identification in a slightly different way (Domon and Aebersold, 2006; Han et al., 2008; Herbert and Johnstone, 2002; Steen and Mann, 2004; Yates, 1998, 2004). Therefore, the goals and budget of the research project must be aligned with the capabilities of the mass spectrometer.

The choice of which mass spectrometer to use will depend on several factors: the type of sample preparation, whether quantitation is expected, the detection of post-translational modification, and the ability to identify as many proteins as possible. First, sample preparation methods may affect the type of instruments used. Some preparation methods are generally associated with how the peptides enter the mass spectrometer. There are two predominant methods for ionization of peptides (Fig. 1): matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) (Yates, 1998). To use MALDI one mixes a protein or peptide sample with a UV-absorbing

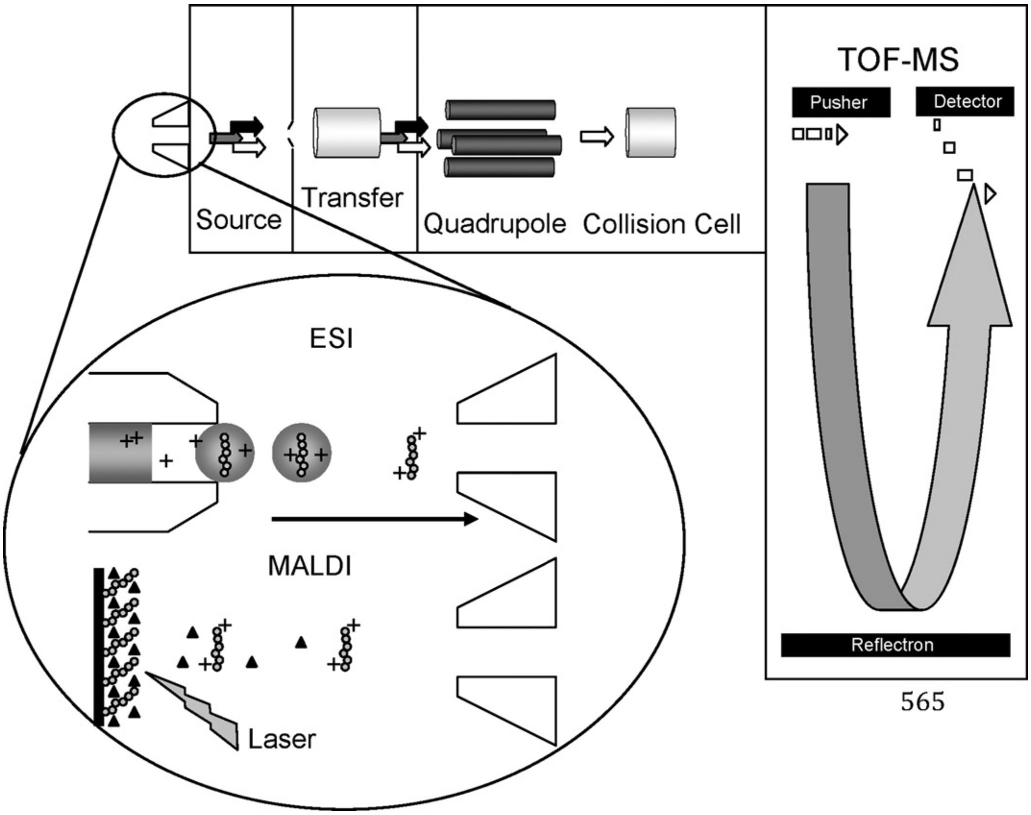

Fig. 1. Tandem mass spectrometer. Ionization of peptides is accomplished by one of two methods, electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI). Charged gas-phase peptides are generated by ESI or MALDI when the acidic HPLC solution containing peptides evaporates or sublimation of peptides occurs when a peptide–crystal matrix is hit with a laser. Ionized gas-phase peptides are then drawn into the mass spectrometer. The qTOF is a tandem mass spectrometer that has a quadrupole mass analyzer in front of a Time-of-Flight mass analyzer. Ionized peptides travel in a constant stream through the instrument to the TOF. There, distinct packets of peptides are pushed orthogonally to their original flight path. A mass spectrum (MS) containing all the peptides in that package can be observed. To sequence a peptide the quadrupole is automatically set to allow only a single peptide to pass. The single peptide is then fragmented in the collision cell prior to entering the TOF, and the peptide's fragments are analyzed.

molecule (matrix) and allows the sample to dry into crystals. Striking the crystals with a UV laser causes rapid heating of the matrix, resulting in sublimation, proton transfer to the protein or peptide, and transfer of the protein or peptide into the mass spectrometer. Multiple laser strikes lead to more ionized sample, and the laser strikes can continue until the sample is consumed. MALDI is typically associated with gel-based forms of protein separation, as specific protein bands or spots from multi-dimensional gel electrophoresis are extracted from the gel and analyzed for content. In contrast, ESI is associated with liquid chromatography methods of protein or peptide isolation. Proteins or peptides in an acidified solvent are ejected from fused silica tubing into a large electrical potential difference between the tubing and the mass spectrometer inlet (Cole, 2000). This forms charged molecules in solvent droplets, where the droplets evaporate prior to the instrument inlet. ESI can be scaled down to solvent flow levels in the nanoliter per minute range in what is referred to as nanoESI. This allows small amounts (picomoles) of sample to be separated by reverse-phase HPLC columns with internal diameter of 75 μm, through tubing with internal diameter of 20–70 μm, out a spray tip as small as 3 μm.

Proteomic experiments with the goal of quantitation of proteins, identification of post-translational modifications, or the identification of as many proteins as possible can be achieved with several different types of mass spectrometers. However, each mass spectrometer has different strengths and weaknesses (Domon and Aebersold, 2006; Han et al., 2008; Herbert and Johnstone, 2002; Steen and Mann, 2004; Yates, 1998, 2004). Most mass spectrometers used in proteomic experiments have multiple mass analyzers in tandem, with the goal of capitalizing on the strengths of each analyzer. Mass analyzers such as quadrupoles (q) and ion traps, are often used in proteomic experiments to hold, filter, and collide ions. Examples of mass analyzers with higher resolving power and accuracy are Time-of-Flight (TOF), Fourier-transform Ion Cyclotron Resonance (FTICR) and Orbitraps; these analyzers allow the sensitivity, accuracy, and resolution necessary to sequence peptides with greater confidence. More sophisticated (and expensive) instruments typically use analyzers with lower resolving power and accuracy, but with the abilities to trap, filter and fragment ions prior to transfer to the high resolving, high accuracy, and high sensitivity analyzers that determine the mass of the ions. Combinations of mass analyzers to form specific instruments (e.g., qTOF is a

Table 1
Strengths of common mass spectrometers used in proteomics.

Instrument	Ion source	Identification	Quantification	Throughput	Detection of modifications
Ion-trap	ESI	+	+++	++	+++
TOF-TOF	MALDI	++	++	+++	+
qTOF	ESI/MALDI	++	+++	++	+
FTICR	ESI/MALDI	+++	++	++	+
Orbitrap	ESI	+++	++	++	++

The strengths of common mass spectrometers used in proteomic experiments are listed. The mass spectrometers are graded as follows: excellent (+++), good (++), and fair (+).

combination of a quadrupole and a Time-of-Flight) have different strengths as far as peptide identification, protein quantitation, sample throughput, and detection of post-translational modifications (Table 1). Optimal results can be expected when experiments are performed with the mass spectrometer best suited for the experiment.

A factor that affects the choice of instrument used, is the cost and availability of the instrument. Mass spectrometers are expensive to purchase, maintain, operate and, therefore, access can be somewhat limited. In addition, operation of a mass spectrometer requires a significant amount of training and expertise, and therefore requires dedicated personnel. Understanding of the physical limits of the instrument is an important factor in knowing what information the experiment can yield.

The time necessary for a mass spectrometer to scan a peptide is an important limitation. Many mass spectrometers have a number of sequential steps that are necessary for sequencing a peptide. For example, a qTOF instrument is a tandem mass spectrometer, meaning that it has two mass analyzers (Fig. 1). The first mass analyzer in a qTOF is a quadrupole (q), followed by the second analyzer the Time-of-Flight (TOF). The first type of scan that an instrument like this collects is a MS scan and this type of scan is detected by the TOF. The MS scan will indicate all the peptides that are entering the instrument at a given time (Fig. 2). If a peptide exceeds a user-defined threshold, the peptide is then sequenced through a process that causes its fragmentation and subsequent determination of all the fragments, this is called the MSMS scan. During a MSMS scan the quadrupole acts as a mass filter, only allowing a single peptide through. That single peptide is then fragmented in a process called collision-induced dissociation (CID). The peptide fragments are then detected by the TOF, and from that pattern, an amino acid sequence is deduced (Fig. 2). The problem with this type of analysis is that it takes time to complete this cycle and during part of this cycle only one peptide is being analyzed when there are potentially hundreds of other peptides entering the instrument at the same time. Depending on instrument settings, the cycle time to sequence 5 peptides can be as long as 25 s.

The problem of instrument cycle time can be illustrated by the following example. If a proteome that consisted of approximately 30,000 proteins was digested and each protein resulted in an average of 10 peptides there would be 300,000 unique peptides. If the peptides were injected onto a reverse-phase HPLC column that is inline with a qTOF, the peptides eluted off the column over a 60-min gradient, and the instrument cycle time to sequence 5 peptides was 25 s, then the number of peptides sequenced would be just over 700 of the 300,000. Mass spectrometers other than the qTOF may have faster cycle time, but some of these instrument sacrifice mass accuracy and/or resolution. Advancements in mass spectrometer are constantly improving mass accuracy, resolution and cycle time. Newer mass spectrometers can sequence multiple peptides at the same time. Although these instruments do not have the scan time issue, the complexity and dynamic range of the proteins in a sample affect the ability to identify proteins. In addition to advances in mass spectrometers, numerous separation techniques have been developed to reduce the number of proteins or peptides in a single sample, spreading them in multiple samples that are individually injected on a mass spectrometer.

3.3. *Complexity and dynamic range*

The Human Proteome Initiative has estimated that the 20,500 human genes could encode up to 1,000,000 different proteins (http://www.expasy.ch/sprot/hpi/). Proteome complexities are in part the result of post-translational modifications and alternate splicing of genes (Boeckmann et al., 2005; Harrison et al., 2002; Kettman et al., 2001).

In addition to the number of different proteins, the proteome is further complicated by protein expression differences that can range as much as 10 orders of magnitude in biological fluids, tissues and cells (Brunet et al., 2003; Panisko et al., 2002; Patterson and Aebersold, 2003). The outcome of this complexity and dynamic range is that high abundance proteins in proteomic surveys mask low abundance proteins of high interest, and require simplification of the proteome for robust mass spectrometry analysis. For example, nearly half of the protein in plasma is albumin, and the top ten proteins in plasma make up nearly 90% of the total protein (Cho, 2007). Various fractionation schemes have been used to enable a more complete identification of proteins (Brunet et al., 2003; Huber et al., 2003; Reinhardt and Lippolis, 2006, 2008; Zolotarjova et al., 2008).

3.4. *Experimental design*

Instrument limitations and the complexity of a sample necessitate experiments to be designed to fractionate and simplify protein samples to enable greater depth of the proteomic discovery. Simplification of a proteome can be achieved by subcellular fractionation, enrichment strategies, chromatography or gel electrophoresis (Stasyk and Huber, 2004). These separation strategies can be used individually or in combination in order to improve detection

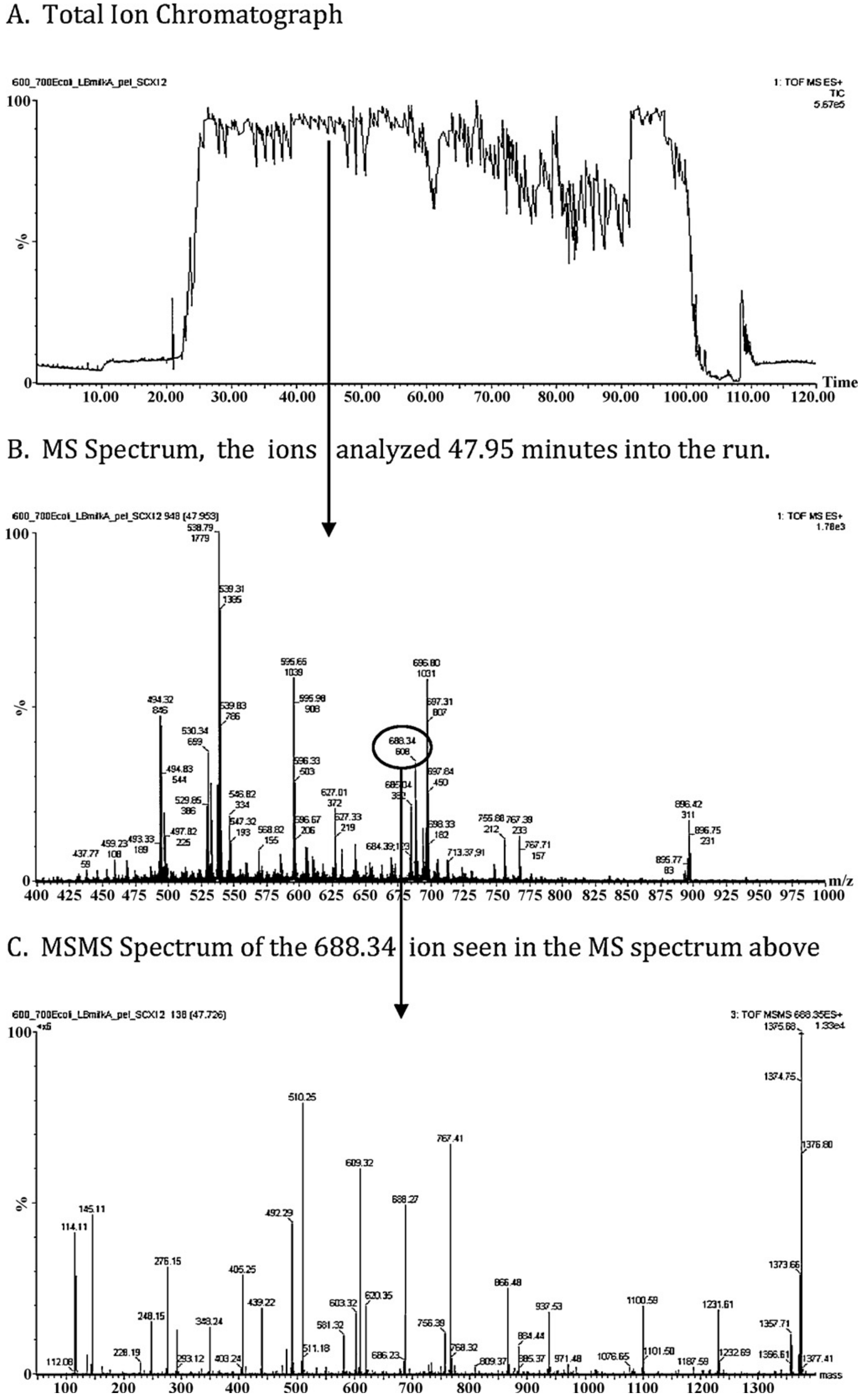

Fig. 2. Mass spectra. This data represents MS and MSMS data from a qTOF mass spectrometer. The top data panel contains a sample MS spectrum. This represents a single 1.5 s scan of the instrument in a 2 h experimental run. At this time point many peptides can be observed. The instrument computer software will select ions with sufficient abundance for subsequent MSMS analysis for sequencing. The bottom panel contains information that led to the sequencing of the 688.3 peptide seen in the top panel.

of low abundance proteins. The choice of which separation strategies to employ will depend on the nature of the sample and the goals of the experiment. For example, a common type of protein separation strategy is referred to as multidimensional protein identification technology (MudPIT), where two different column materials are packed into the same HPLC column (Washburn et al., 2001). In this type of experiment a strong cation exchange (SCX) column matrix is packed next to reverse-phase (RP) matrix. Peptides are injected onto the column and bind to the SCX material. Sequential solutions with different concentrations of salt will free a group of specific peptides from the SCX material and bind to the subsequent RP material. Between each increase in salt concentration, a RP gradient is run, which does not affect the peptides still bound to the SCX column, but will remove peptides bound to the RP column into the mass spectrometer for analysis. Using this technique investigators were able to identify 1484 proteins from *Saccharomyces cerevisiae*, whereas previous experiments using an alternate separation technique yielded just fewer than 300 protein identifications (Washburn et al., 2001).

However, ~1500 proteins identified by mass spectrometry are still well short of the tens of thousands of proteins in *S. cerevisiae*. One alternative is to isolate subcellular components for proteomic analysis, which would likely have a reduced protein complexity. An example of this approach can be seen in the proteomic analysis of bovine milk fat globule membranes (MFGM) (Reinhardt and Lippolis, 2006). MFGM represent the apical membrane of secretory mammary epithelial cells and are easily obtained in milk. As such MFGM contain the proteins necessary for transport of milk, milk fats and minerals such as calcium. However, proteomics of the MFGM is complicated by the fact that between 4 and 8 proteins comprise 60% of the total proteins by weight and 1 protein, butyrophilin, accounts for 30–40% of the total protein by itself (Mather, 2000).

Biological samples can contain a few proteins that make up the majority of the total protein content of the sample; MFGM and plasma are two such examples of this phenomenon. As stated earlier nearly half of the protein in plasma is albumin, and the top ten proteins in plasma make up nearly 90% of the total protein (Cho, 2007). The problem is that most mass spectrometers pick peptides to sequence on the basis of signal intensity; therefore the peptides with the highest signal intensity are more likely to be sequenced. In addition, peptides with higher signal intensity will be more likely to yield an unambiguous spectrum to deduce the sequence. Various solutions to this problem have been proposed, from depletion of abundant proteins to alternate separation strategies. Using a method to deplete abundant proteins, it has been possible to extend the detection of plasma proteins from μg/ml to ng/ml (Anderson and Hunter, 2006).

There are currently two major strategies to deplete overly abundant proteins (Polaskova et al., 2010). Antibody cocktails are commercially available that contain antibodies against the 12 highest abundance human plasma proteins. Thus, specific removal of the most abundant human plasma proteins has yielded greater depth of the proteomic dataset. However, it has also been observed that depletion of abundant proteins can result in the non-specific loss of low abundance proteins (Granger et al., 2005). A second method of depletion uses multiple affinity removal columns, which have random hexapeptides bound to a matrix (Thulasiraman et al., 2005). These hexapeptides provide non-covalent binding sites to capture proteins. Because each amount of each hexapeptide is limited, high abundance proteins are reduced at the same time low abundance proteins are concentrated. The goal of both of these techniques is to eliminate the signal suppression by high-abundance proteins and thus amplify the signal of low-abundance proteins. This later method precludes protein quantitation proteomics.

In addition to depletion strategies are those methods that use antibodies and metal ions as a means to enrich for a specific type of protein. For example, antibodies specific for a class of molecules called major histocompatibility complex (MHC) have been used to isolate these molecules away from other cellular proteins. A MHC molecule binds to a variety of peptide fragments both from proteins normally found in a cell and importantly from pathogens. The MHC–peptide complex is then present on the cell surface for detection by the immune system. MHC molecules and their associated peptides can be precipitated using monoclonal antibodies. The peptides are then separated from the MHC molecule using size exclusion filtration, and sequenced using mass spectrometry (Hunt et al., 1992; Lippolis et al., 2002). Examples of this type of proteomic approach include the determination of the nature of autoantigens that may be involved in the autoimmune disease type 1 diabetes (Nepom et al., 2001) and the nature of antigens involved in the response to diseases such as melanoma (Cox et al., 1994). In addition to the specific isolation of a molecule, antibodies can be used to isolate a class of molecules. Both antibodies and metal ions (immobilized metal ion affinity chromatography [IMAC]) have been used to enrich for phosphorylated proteins (Ptacek and Snyder, 2006). It is estimated that 30% of cellular proteins are phosphorylated, and phosphorylation often acts as an on/off switch for the protein's function. The isolation and subsequent identification of the phosphoproteome is an important new area of discovery.

Non-depletion strategies can also be used, but they also have their limitations. In the example of MFGM where 1 protein, butyrophilin, represents 30–40% of the total protein, investigators chose to isolate the butyrophilin away from the other proteins by SDS-PAGE (Reinhardt and Lippolis, 2006). In short, total MFGM proteins were loaded on a gel and proteins separated. The gel was cut into 37 gel slices, isolating various protein bands. The proteins were then digested and extracted from the gel for analysis by the mass spectrometer. Despite the clear butyrophilin band in the gel (Fig. 3), butyrophilin peptides were detected in every part of the gel. In subsequent analysis of the data we did show that in the area of the butyrophilin band, over 90% of the MSMS spectra in that slice of the gel are identified as butyrophilin peptides and in area distant to the butyrophilin bands a large number of MSMS spectra are identified as butyrophilin (Fig. 3). This shows that a single very abundant protein can cause major problems for

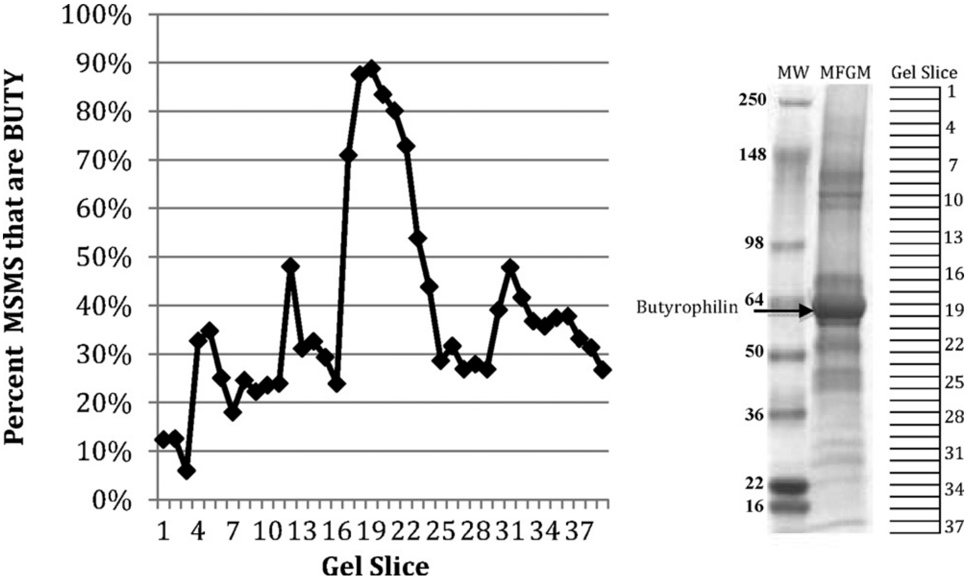

Fig. 3. Gel electrophoresis of proteome. Milk fat globule membrane proteome was separated by one-dimensional gel electrophoresis. The gel was sliced (37 times) and proteins were digested with trypsin in the gel. Peptides were extracted from the gel and the peptides from each gel slice were run on a HPLC column coupled to a tandem mass spectrometer (qTOF). The mass spectrometry data was analyzed using Mascot (Matrix Science) to predict protein matches. All MSMS spectra that were matched to butyrophilin were quantitated for each gel slice and graphed above.

proteomic analysis. In addition, separation techniques such as gel electrophoresis do not have the resolving power to mitigate the interference of an overly abundant protein.

3.5. Data validation

Proteomic experiments generate a large quantity of mass spectrometry data. A typical MudPIT proteomic experiment can yield tens of thousands of MSMS spectra, each requiring the determination of the peptide sequence. Gigabytes of information must be first distilled down to peptide sequences and those sequences matched to proteins, in a process that is termed computational proteomics (reviewed by Colinge and Bennett, 2007; McHugh and Arthur, 2008). Gone are the days of manually determining a peptide sequence from the MSMS spectrum. The number of tools available for computational proteomics has expanded greatly in the last decade. The pace that the field of proteomics will advance is dependent on the development of hardware, software, and data management tools (Table 2).

The first step in understanding the data obtained from the mass spectrometer is to align the data with proteins in the databases. Despite improved software, the continuing challenge for researchers who use proteomics as a tool is how to interpret the data and how much confidence one can have in the proteins identified. How many unique peptides are required to identify a protein? What is the false discovery rate? How are closely related proteins distinguished? These questions are presently a matter of debate in the proteomics field. However, these are the questions that one must address when writing and reading a proteomics paper (Carr et al., 2004; Quadroni et al., 2004; Wilkins et al., 2006).

There are many software packages that take different approaches to transform mass spectrometry data into pro-

Table 2
Various proteomic software packages.

Name	Function	Website
Mascot	Peptide and protein identification/quantitation	www.matrixscience.com
Sequest	Peptide and protein identification/quantitation	www.thermo.com
ProteinProphet	Validation of protein identification	http://proteinprophet.sourceforge.net/
X!Tandem	Peptide matching with MSMS spectra	www.thegpm.org
Scaffold	Peptide and protein identification and data organization	www.proteomesoftware.com
Trans-proteomic pipeline	Peptide and protein identification and quantitation. Tools for data visualization	http://tools.proteomecenter.org/wiki
The Gene Ontology	Protein functional database	www.geneontology.org
Kyoto Encyclopedia of Genes and Genomes	Protein functional database	www.genome.jp/kegg
PEAKS	De novo search program for peptide and protein identification	www.bioinformaticssolutions.com

A brief list of common proteomic software packages.

tein identification and expression data. In a study that compared four peptide identification algorithms to identify peptides from a complex protein sample, 608 peptides were identified by one or more search algorithm. Of the 608 peptides identified, only 335 peptides were correctly identified by all four search programs, and a range from 0 to 46 peptides were correctly identified by only one search program (Kapp et al., 2005). This data shows that there can be a significant number of differences in MSMS spectra interpretations of a dataset by various software packages, and that the use of multiple software packages may add to the confidence in the identification of the peptides.

Evaluation of mass spectrometry data is typically accomplished by the major software packages that interpret the dataset. Most of these software packages report a score on which the confidence of the mass spectrum solution is based. Based on the number of peptides and their scores, a protein score is calculated that can indicate the confidence that one can have in the identifications of proteins that exceed a threshold score. As discussed above, the various software packages can identify different subsets of proteins from the mass spectrometry data. Therefore, those that are identified by multiple software packages would have a higher confidence than those that were identified by a single software package.

In order to better determine the error rate of protein identification, methods have been developed to determine what is called the false discovery rate. In this method MSMS spectra are searched against a protein database of choice followed by a search against the same protein database where the sequences have been reversed, randomized, or shuffled (decoy database) (Elias and Gygi, 2007). The expectation is that there will not be any true matches to the decoy database. Therefore, any matches to the decoy database are defined as false positive matches. The false discovery rate is calculated by calculating the number of false positive matches divided by the sum of the number of true positive matches plus the number of false positive matches. This information can help the investigator alter the significance thresholds to optimize the data to limit the reporting of false positive protein identifications.

4. Promise: what will proteomics be able to do?

Despite the limitations that the field of proteomics currently has to deal with, it has become an extremely important tool in biological sciences. The first unique advantage of this technology is the fact that a fairly large number of proteins can be identified and quantitated at one time, without any prior knowledge that any specific protein might exist in a sample. Analyzing a proteomic dataset can often lead to surprising results, and the unexpected may be the most interesting observation. In fact, most shotgun proteomic experiments are not typical "hypothesis driven" experiments, but may be better considered experiments designed to find a hypothesis. In these experiments hundred of proteins can be identified whose expression is altered by a defined experimental condition. Some changes in protein expression may be expected and even well characterized. However, some may be unexpected or unknown and lead to new hypothesis for the connections between protein expression and cellular processes. For example, comparing the protein expression from *E. coli* grown in milk versus laboratory media, it is quite expected that proteins, such as beta-galactosidase, involved in lactose utilization would be up regulated in those bacteria grown in milk (Lippolis et al., 2009). In contrast, the protein S-rebosylhomocysteine lyase or the LuxS gene product was up regulated in bacteria grown in milk. This enzyme is critical for the synthesis of a bacterial hormone like compound called autoinducer-2 (AI-2) (Sperandio et al., 2003). AI-2 is involved in the regulation of hundreds of genes, many of which are virulence genes (DeLisa et al., 2001). Whereas the observation of beta-galactosidase perhaps was expected, the observations of the LuxS gene product was not and has lead to the generation of hypothesis from this unexpected result.

Other protein detection methodologies, such as western blots, rely on the availability of quality antibodies. Animal scientists are fully aware of the limited number of antibodies available to them when compared to those specific for human or rodent proteins. It is this independence from antibodies that makes proteomic experiments especially attractive to those wanting to identify proteins in domestic animals. Furthermore, the breadth of data generated by methodologies using antibodies would be less than a sweeping proteomic survey.

Although mRNA quantitation methods are very powerful tools there is the limitation that they do not measure the functional end product, which is the protein. Numerous studies have shown examples of a lack of correlation between mRNA and protein abundance (Griffin et al., 2002; Gygi et al., 1999; Ideker et al., 2001). For example, changing the carbon source for yeast resulted in a 500-fold increase in mRNA for a gene involved in sugar metabolism, whereas the corresponding protein only increased 10-fold. In addition, some genes showed no change in mRNA levels but showed significant increases in protein levels (Griffin et al., 2002). These examples of a lack of correlation highlight the importance of linking mRNA expression results with subsequent proteomic studies. In addition, mRNA expression experiments cannot detect protein post-translational modifications.

Proteomic experiments are also ideally suited for detection and identification of post-translational modifications on proteins. Antibody based detection of phosphorylated proteins are limited to detection of a whole class of phosphoproteins (e.g., proteins with a phospho-serine) or a specific phosphorylation event on a specific protein. Proteomics offers the as of yet not fully realized promise of analysis of an entire phosphoproteome. The ability to observe the effect of a stimulus of the entire phosphoproteome with time would be a very powerful tool in understanding the complex and integrated intracellular signaling mechanisms.

The ultimate goal of proteomics is to detect and quantify the tens of thousands of proteins that constitute a proteome. To be able to observe the entire cellular picture of protein expression, location, interaction, and modification under different experimental conditions would aid in the understanding of the molecular mechanism critical for cellular functions. This understanding of cellular functions would be the basis for rational therapeutic designs to target pathogens and correct disease conditions.

Conflict of interest

Authors have no financial or other relationships that would inappropriately influence this work.

References

Aebersold, R.H., Mann, M., 2003. Mass spectrometry-based proteomics. Nature 422, 198–207.

Anderson, L., Hunter, C.L., 2006. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol. Cell. Proteomics 5, 573–588.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B., 2007. Quantitative mass spectrometry in proteomics: a critical review. Anal. Bioanal. Chem. 389, 1017–1031.

Boeckmann, B., Blatter, M.-C., Famiglietti, L., Hinz, U., Lane, L., Roechert, B., Bairoch, A., 2005. Protein variety and functional diversity: Swiss-Prot annotation in its biological context. C. R. Biol. 328, 882–899.

Brunet, S., Thibault, P., Gagnon, E., Kearney, P., Bergeron, J.J.M., Desjardins, M., 2003. Organelle proteomics: looking at less to see more. Trends Cell Biol. 13, 629–638.

Carr, S.A., Aebersold, R.H., Baldwin, M., Burlingame, A.L., Clauser, K., Nesvizhskii, A., Working Group on Publication Guidelines for Peptide and Protein Identification Data, 2004. The need for guidelines in publication of peptide and protein identification data. Mol. Cell. Proteomics 3, 531–533.

Cho, W.C.S., 2007. Proteomics technologies and challenges. Genomics Proteomics Bioinform. 5, 77–85.

Cole, R.B., 2000. Some tenets pertaining to electrospray ionization mass spectrometry. J. Mass Spectrom. 35, 763–772.

Colinge, J., Bennett, K.L., 2007. Introduction to computational proteomics. PLoS Comput. Biol. 3, e114.

Cox, A.L., Skipper, J., Chen, Y., Henderson, R.A., Darrow, T.L., Shabanowitz, J., Engelhard, V.H., Hunt, D.F., Slingluff, C.L., 1994. Identification of a peptide recognized by five melanoma-specific human cytotoxic T cell lines. Science 264, 716–719.

Cravatt, B.F., Simon, G.M., Yates, J.R., 2007. The biological impact of mass-spectrometry-based proteomics. Nature 450, 991–1000.

DeLisa, M.P., Wu, C.F., Wang, L., Valdes, J.J., Bentley, W.E., 2001. DNA microarray-based identification of genes controlled by autoinducer 2-stimulated quorum sensing in *Escherichia coli*. J. Bacteriol. 183, 5239–5247.

Domon, B., Aebersold, R.H., 2006. Mass spectrometry and protein analysis. Science 312, 212–217.

Elias, J.E., Gygi, S.P., 2007. Target–decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods 4, 207–214.

Gingras, A.-C., Gstaiger, M., Raught, B., Aebersold, R.H., 2007. Analysis of protein complexes using mass spectrometry. Nat. Rev. Mol. Cell Biol. 8, 645–654.

Granger, J., Siddiqui, J., Copeland, S., Remick, D., 2005. Albumin depletion of human plasma also removes low abundance proteins including the cytokines. Proteomics 5, 4713–4718.

Griffin, T.J., Gygi, S.P., Ideker, T., Rist, B., Eng, J.K., Hood, L., Aebersold, R.H., 2002. Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. Mol. Cell. Proteomics 1, 323–333.

Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R.H., 1999. Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. 19, 1720–1730.

Han, X., Aslanian, A., Yates, J.R., 2008. Mass spectrometry for proteomics. Curr. Opin. Chem. Biol. 12, 483–490.

Harrison, P.M., Kumar, A., Lang, N., Snyder, M., Gerstein, M., 2002. A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Res. 30, 1083–1090.

Herbert, C.B., Johnstone, R.A.W., 2002. Quadrupole Ion Optics. Mass Spectrometry Basics, pp. 183–188.

Huber, L.A., Pfaller, K., Vietor, I., 2003. Organelle proteomics: implications for subcellular fractionation in proteomics. Circ. Res. 92, 962–968.

Hunt, D.F., Henderson, R.A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A.L., Appella, E., Engelhard, V.H., 1992. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. Science 255, 1261–1263.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R.H., Hood, L., 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292, 929–934.

Kapp, E.A., Schütz, F., Connolly, L.M., Chakel, J.A., Meza, J.E., Miller, C.A., Fenyo, D., Eng, J.K., Adkins, J.N., Omenn, G.S., Simpson, R.J., 2005. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 5, 3475–3490.

Kettman, J.R., Frey, J.R., Lefkovits, I., 2001. Proteome, transcriptome and genome: top down or bottom up analysis? Biomol. Eng. 18, 207–212.

Klebba, P.E., 2003. Three paradoxes of ferric enterobactin uptake. Front. Biosci. 8, s1422–s1436.

Kornalijnslijper, J.E., Daemen, A.J.J.M., van Werven, T., Niewold, T.A., Rutten, V.P.M.G., Noordhuizen-Stassen, E.N., 2004. Bacterial growth during the early phase of infection determines the severity of experimental *Escherichia coli* mastitis in dairy cows. Vet. Microbiol. 101, 177–186.

Legrand, D., Elass, E., Carpentier, M., Mazurier, J., 2005. Lactoferrin: a modulator of immune and inflammatory responses. Cell. Mol. Life Sci. 62, 2549–2559.

Lippolis, J.D., Bayles, D.O., Reinhardt, T.A., 2009. Proteomic changes in *Escherichia coli* when grown in fresh milk versus laboratory media. J. Proteome Res. 8, 149–158.

Lippolis, J.D., Reinhardt, T.A., 2005. Proteomic survey of bovine neutrophils. Vet. Immunol. Immunopathol. 103, 53–65.

Lippolis, J.D., Reinhardt, T.A., 2008. Centennial paper: proteomics in animal science. J. Anim. Sci. 86, 2430–2441.

Lippolis, J.D., White, F.M., Marto, J.A., Luckey, C.J., Bullock, T.N.J., Shabanowitz, J., Hunt, D.F., Engelhard, V.H., 2002. Analysis of MHC class II antigen processing by quantitation of peptides that constitute nested sets. J. Immunol. 169, 5089–5097.

Martin, S.E., Shabanowitz, J., Hunt, D.F., Marto, J.A., 2000. Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. Anal. Chem. 72, 4266–4274.

Mather, I.H., 2000. A review and proposed nomenclature for major proteins of the milk–fat globule membrane. J. Dairy Sci. 83, 203–247.

McHugh, L., Arthur, J.W., 2008. Computational methods for protein identification from mass spectrometry data. PLoS Comput. Biol. 4, e12.

Moyer, S.C., Budnik, B.A., Pittman, J.L., Costello, C.E., O'Connor, P.B., 2003. Attomole peptide analysis by high-pressure matrix-assisted laser desorption/ionization Fourier transform mass spectrometry. Anal. Chem. 75, 6449–6454.

Nepom, G.T., Lippolis, J.D., White, F.M., Masewicz, S., Marto, J.A., Herman, A., Luckey, C.J., Falk, B., Shabanowitz, J., Hunt, D.F., Engelhard, V.H., Nepom, B.S., 2001. Identification and modulation of a naturally processed T cell epitope from the diabetes-associated autoantigen human glutamic acid decarboxylase 65 (hGAD65). Proc. Natl. Acad. Sci. U.S.A. 98, 1763–1768.

Ong, S.-E., Mann, M., 2005. Mass spectrometry-based proteomics turns quantitative. Nat. Chem. Biol. 1, 252–262.

Panisko, E.A., Conrads, T.P., Goshe, M.B., Veenstra, T.D., 2002. The postgenomic age: characterization of proteomes. Exp. Hematol. 30, 97–107.

Patterson, S.D., Aebersold, R.H., 2003. Proteomics: the first decade and beyond. Nat. Genet. 33 (Suppl.), 311–323.

Polaskova, V., Kapur, A., Khan, A., Molloy, M.P., Baker, M.S., 2010. High-abundance protein depletion: comparison of methods for human plasma biomarker discovery. Electrophoresis 31, 471–482.

Ptacek, J., Snyder, M., 2006. Charging it up: global analysis of protein phosphorylation. Trends Genet. 22, 545–554.

Quadroni, M., Ducret, A., Stöcklin, R., 2004. Quantify this! Report on a round table discussion on quantitative mass spectrometry in proteomics. Proteomics 4, 2211–2215.

Reinhardt, T.A., Lippolis, J.D., 2006. Bovine milk fat globule membrane proteome. J. Dairy Res. 73, 406–416.

Reinhardt, T.A., Lippolis, J.D., 2008. Developmental changes in the milk fat globule membrane proteome during the transition from colostrum to milk. J. Dairy Sci. 91, 2307–2318.

Sperandio, V., Torres, A.G., Jarvis, B., Nataro, J.P., Kaper, J.B., 2003. Bacteria–host communication: the language of hormones. Proc. Natl. Acad. Sci. U.S.A. 100, 8951–8956.

Stasyk, T., Huber, L.A., 2004. Zooming in: fractionation strategies in proteomics. Proteomics 4, 3704–3716.

Steen, H., Mann, M., 2004. The ABC's (and XYZ's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. 5, 699–711.

Takemura, K., Hogan, J.S., Lin, J., Smith, K.L., 2002. Efficacy of immunization with ferric citrate receptor FecA from *Escherichia coli* on induced coliform mastitis. J. Dairy Sci. 85, 774–781.

Thulasiraman, V., Lin, S., Gheorghiu, L., Lathrop, J., Lomas, L., Hammond, D., Boschetti, E., 2005. Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands. Electrophoresis 26, 3561–3571.

Washburn, M.P., Wolters, D., Yates, J.R., 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. 19, 242–247.

Wilkins, M.R., Appel, R.D., Van Eyk, J.E., Chung, M.C.M., Görg, A., Hecker, M., Huber, L.A., Langen, H., Link, A.J., Paik, Y.-K., Patterson, S.D., Pennington, S.R., Rabilloud, T., Simpson, R.J., Weiss, W., Dunn, M.J., 2006. Guidelines for the next 10 years of proteomics. Proteomics 6, 4–8.

Witze, E.S., Old, W.M., Resing, K.A., Ahn, N.G., 2007. Mapping protein post-translational modifications with mass spectrometry. Nat. Methods 4, 798–806.

Yates, J.R., 1998. Mass spectrometry and the age of the proteome. J. Mass Spectrom. 33, 1–19.

Yates, J.R., 2004. Mass spectral analysis in proteomics. Annu. Rev. Biophys. Biomol. Struct. 33, 297–316.

Zolotarjova, N., Mrozinski, P., Chen, H., Martosella, J., 2008. Combination of affinity depletion of abundant proteins and reversed-phase fractionation in proteomic analysis of human plasma/serum. J. Chromatogr. A 1189, 332–338.